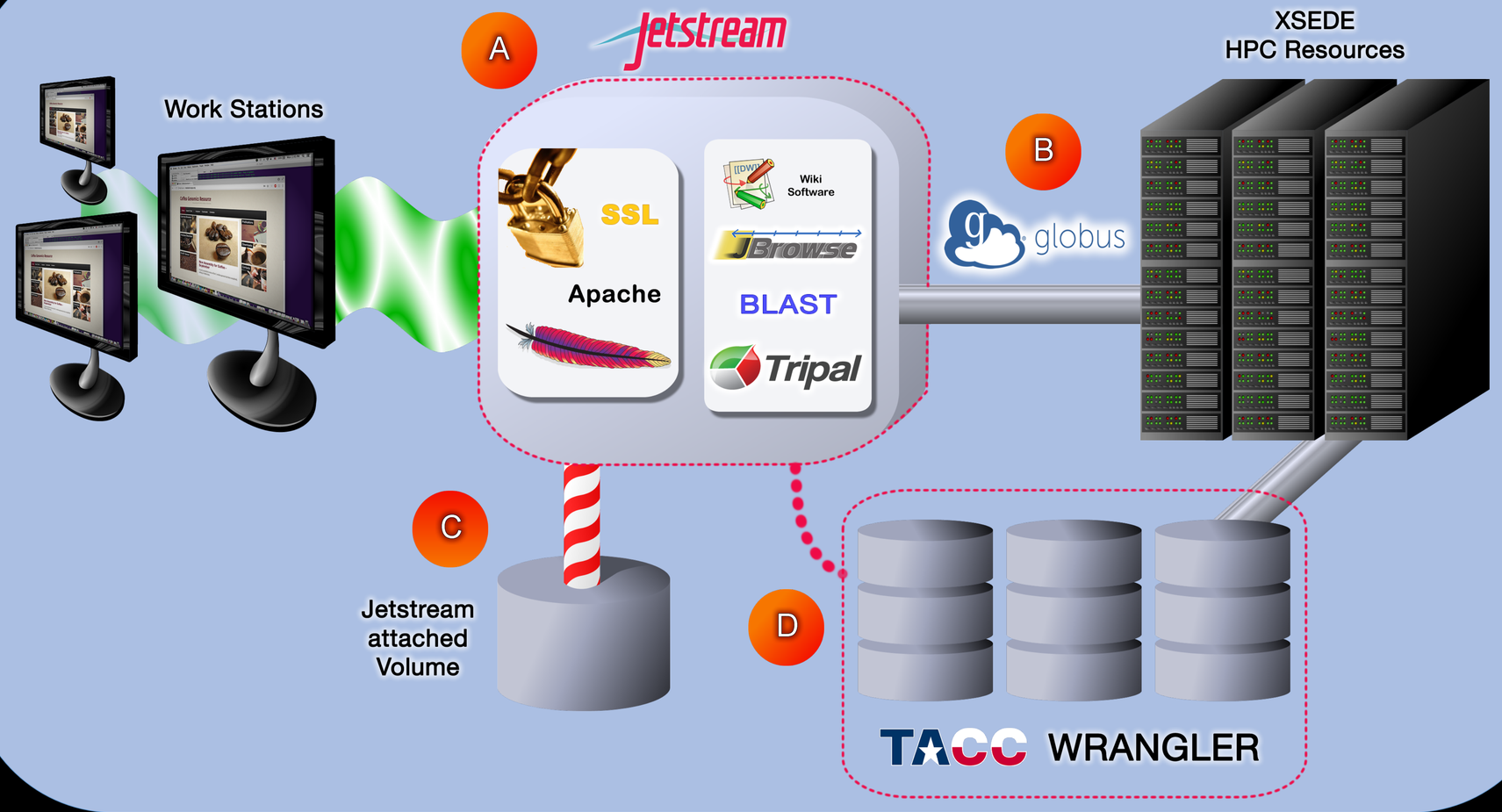
Solving the Challenges of Complex Genome Analysis Collaborations On-line Using XSEDE Resources

Carrie Ganote[a], Sheri Sanders[a], Le-Shin Wu[a,b], Thomas G. Doak[a], Keithanne Mockaitis[a,c]

[a]Pervasive Technology Institute, Indiana University, Bloomington IN USA
[b]Eli Lilly and Company, Indianapolis IN USA
[c]Department of Biology, Indiana University, Bloomington IN USA

Abstract
Whole genome reference and functional genomics projects most often benefit from a diversity of expertise and the integrated contributions of multiple research groups. The importance of web-enabled data sharing and open access software to progress in genome research is indisputable. Disseminating genome resource information through internet-based resources, especially customizing and encouraging the optimal use of analysis tools to serve a specific research community, often lags behind data generation. This lag can inhibit biological interpretations and downstream experimentation, much of which should be undertaken before a genome resource project is completed. File systems for storing data and analysis outputs of today's project standards must be large and secure, and must offer sustained access to fulfill the hosting requirements an active and dispersed research community needs. Users must interact with a computing resource powerful enough to get their jobs done, and sites must be expandable and flexible to accommodate a growing demand for intra-genera comparisons and pan-genomics of a species. We have been using NSF XSEDE computational resources including Jetstream, along with the High Performance Computing systems of Indiana University to meet these challenges for a variety of collaborative plant genomics studies. Currently these efforts are impacting metabolic gene discovery in the tetraploid *Arachis hypogaea* and whole genomic reference studies of the tetraploid *Coffea arabica* and its diploid progenitors. Here we show practical examples of XSEDE resource use and development that may benefit other genomic research groups seeking to increase the effectiveness of their computing and collaboration.



File System Challenges
Data management is a difficult, perennial, and ubiquitous problem in genomics projects due to:

-ever-growing volumes of genomic data
-diversity of analysis approaches & intermediate files
-institutional policies, quotas or time limits governing the data storage
-potential for security breaches

Finding enough storage and memory can be a challenge. Smaller research facilities in particular quickly outgrow their computing resources. Finding alternative options for data storage and use is especially important, and often overlooked, in the design and utilization of web servers using virtual machines (VMs). Individual VMs are often allocated space that is insufficient for most genomics projects, as reference sequence files, gff files, bam files, bed files, etc. tend to be larger than most types of files commonly populating web sites; genome visualization software such as JBrowse will create and utilize a huge number of small files when preparing genomic alignments tracks which present additional challenges for copying, accessing and hosting data for display. Mounting the research file system onto the VM serving the research web site can allow for security holes.

File System Solutions
We now use block storage on the XSEDE[1] cloud computing resource Jetstream (Volumes), with offers up to a total of 500 GB[2]. See Figure (C).
Jetstream Volumes works for a single VM, but become problematic when expansions are needed, such as adding new .gff files to a JBrowse configuration, as Volumes can only be attached to one running VM at a time.

JBrowse requires memory and CPU power to generate display-ready files from reference files. With the priority to use compute cycles frugally, our Jetstream VM runs web functions at the lowest resource drain possible. In order to expand file sets, we need to shut the VM down and create another with higher capacity, or else run the reference preparation on another machine and copy the files over. Copying the large numbers of small files generated by JBrowse is time consuming, so the latter is a less attractive option. Globus (see Figure B) can make this faster, but we found it best to share a file system directly between two machines.

Our allocation on XSEDE resource Wrangler allows us to mount the Wrangler file system onto Jetstream VMs using Network File System (NFS) – see Figure (D). We found this to be the best option for sharing data between VMs. We also benefit from being able to run the generating scripts on Wrangler, in lieu of creating a compute-cycle expensive VM. The performance of serving the resulting browser files over NFS to the web is being tested, and more solutions will be explored as needs arise.

Stable long term hosting
An often-overlooked aspect of data dissemination and ongoing collaborations is a long-term web presence for the project. Web hosting must be secure, funded, and maintained for the life of the project. To serve this need we utilize:
1) the Intelligent Infrastructure Virtual Systems (II), a fee-for-service resource at Indiana University that offers VMs with high availability, and
2) Jetstream, the XSEDE cloud system tailored for research use and needs[2].

We tested the XSEDE resource Wrangler as an alternative solution to the web hosting issue. Wrangler offers a very generous amount of storage coupled with tightly integrated compute nodes. It is intended for data-intensive computing, and we tested it in hopes that it would be ideal for sending BLAST jobs from Tripal to the Wrangler compute nodes. Disadvantages however were the lack of root privileges, lack of access to the web server from Tripal, and difficulty sharing files using UNIX groups. Ultimately we abandoned Wrangler as our sole web and file system resource.

Jetstream benefits include VMs that can be created, destroyed, and suspended as needed – see Figure (A). Researchers are charged only for time used. The "charge" for time on Jetstream follows the XSEDE model, in which one is awarded computational hours according to a straightforward proposal submitted to XSEDE. Following a 'startup' allocation, a more detailed, full proposal justifying the amount of time needed for a project's lifetime is submitted.

Collaborating through Tripal
Tripal [3,4] is an extension of the content management system Drupal 5 that optimizes management of genomic data by using the GMOD community database schema Chado[6]. The Tripal platform is attractive in collaboration settings as it allows relatively easy setup of a central hub of data, visualization, analysis, and discussion. The benefits of the platform are:

-Ease of use
-Ease of customization
-Content manager with a well-established graphic user interface (GUI)
-Automated web page creation (point-and-click)
-Preconfigured modules allow for annotation of analyses
-Organized record keeping of analyses and discussions
-Secure user vetting and page permission configurations

Advantages for Security: The security handling lets us easily set up individual users and assign them a highly customizable level of access. For instance, we can limit to administrators the setting up of new JBrowse instances via our GUI (see below), but allow any authenticated user to add tracks to existing JBrowse instances. This provides us with site management that is much more advanced than our previous security implementations using .htaccess, which offered limited granularity of permission and non-automated user setup.

Advantages in File Serving: With the above fine-grain security handling in place, it is painless to provide access to FTP sites, file downloads, etc. to authorized users, while publishing prepared data for the public.

Expertise
NCGAS, the National Center for Genome Analysis Support, is an NSF-funded center with the mission of assisting any NSF-funded genomics project by providing:

-access to computational resources
-curated sets for genomics applications
-expert consultation

Our strengths reflect a well-balanced team with a wide variety of talents. We offer expertise in high performance computing, web development and Biology.

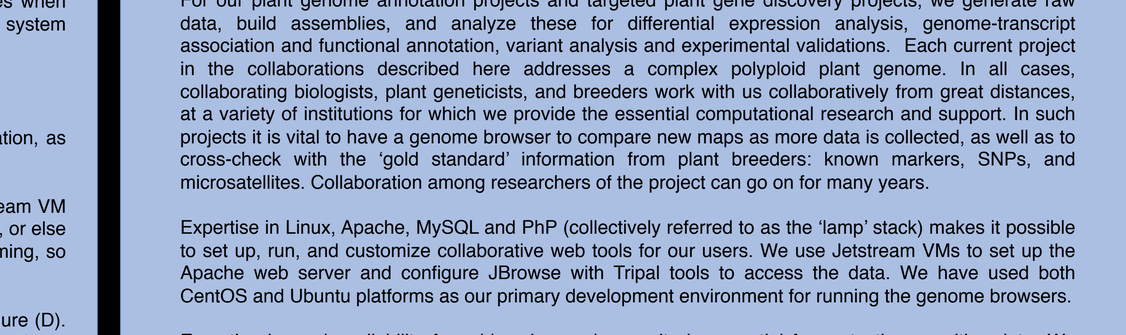
For our plant genome annotation projects and targeted plant gene discovery projects, we generate raw data, build assemblies, and analyze these for differential expression analysis, genome-transcript association and functional annotation, variant analysis and experimental validations. Each current project in the collaborations described here addresses a complex polyploid plant genome. In all cases, collaborating biologists, plant geneticists, and breeders work with us collaboratively from great distances, at a variety of institutions for which we provide the essential computational research and support. In such projects it is vital to have a genome browser to compare new maps as more data is collected, as well as to cross-check with the 'gold standard' information from plant breeders: known markers, SNPs, and microsatellites. Collaboration among researchers of the project can go on for many years.

Expertise in Linux, Apache, MySQL and PhP (collectively referred to as the 'lamp' stack) makes it possible to set up, run, and customize collaborative web tools for our users. We use Jetstream VMs to set up the Apache web server and configure JBrowse with Tripal tools to access the data. We have used both CentOS and Ubuntu platforms as our primary development environment for running the genome browsers.

Expertise in, and availability for addressing, web security is essential for protecting sensitive data. We prioritize security, including maintenance of Firewalls and use of SSL encryption.

The Application Programming Interface (API) of Jetstream allows users to communicate directly to OpenStack software from the command line or from inside a script. This opens up many possibilities for advanced customizations, such as creation of a hundred VMs at once or configuring IP addresses precisely. Our API expertise with the OpenStack system gives us an edge when dealing with cloud resources.

Our expertise with Tripal tools has developed the initial setup of a centralized site for collaboration to be much less command line dependent[7]. Our new module allows for the on-demand spawning of new JBrowse references and tracks for authorized users, removing the need for administrators to manage visualization for an entire community. We have also modified the BLAST_UI module to interface with these and other JBrowse instances, removing the need for users to request the addition of new tracks or to learn how to modify URLs in JBrowse to serve blast results. These tools were developed recently to serve our NSF-funded Coffee genomics project and the USDA-funded Peanut transcriptomics project, and are now made available to the community as a whole. Both of these tools and other standards (BLAST server, job daemon, security management) are implemented on a ready to use, ready to launch VM publicly available on Jetstream[2], which allows communities to start with minimal initial knowledge and command line use.



Summary

NCGAS provides the skills and resources needed for contemporary genomics research. We have overcome challenges of long-distance collaboration for a number of genome reference building and genomics discovery projects. Today we leverage US national resources such as XSEDE, and have created links between resources in novel ways to take advantage of the best features of each.

Our collaboration tools of choice are Globus for data sharing, Genome browsers (primarily JBrowse) tied to web pages using Tripal. We have optimized resource use to benefit our collaborations with file systems that fit our needs, software development to make tasks easier, tape archives for data back up, and hosts capable of long-term, economical, and reliable web serving.

References
[1]https://portal.xsede.org
[2]See Papudeshi B, PAG XXVI C31: Wednesday, 12:00-12:15 pm and https://iujetstream.atlassian.net/wiki/spaces/JWT/pages/32899113/Volumes
[3]Ficklin SP, Sanderson LA, Cheng CH, Staton ME, Lee T, Cho IH, Jung S, Bett KE, Main D. 2011. Tripal: a10.1093/database/bat075 construction toolkit for online genome databases. Database (Oxford). doi: 10.1093/database/bar044
[4]Sanderson LA, Ficklin SP, Cheng CH, Jung S, Feltus FA, Bett KE, Main D. 2013 Tripal v1.1: a standards-based toolkit for construction of online genetic and genomic databases. Database (Oxford). doi:
[5]http://drupal.org
[6] http://gmod.org
[7]Sanders S, PAGXXVI W1030, Sunday 5:30-5:48 pm

CG, SS and KM were supported by NSF Grant IOS-1444893 (KM); SS and KM were supported by USDA-NIFA Grant 2014-67014-21701 (KM); CG, SS L-SW and TD were supported by NSF Grants DBI-1458641 and ABI-1062432, *National Center for Genome Analysis Support, a collaboration between Indiana University and the Pittsburgh Supercomputing Center*.